On Estimating the Phase of a Periodic Waveform in Additive Gaussian Noise — Part I

L. L. Rauch
Communications Systems Research Section

Motivated by recent advances in technology, a new look is taken at the problem of estimating the phase of a periodic waveform in additive gaussian noise. The maximum a posteriori probability criterion with signal space interpretation is used to obtain the structures of optimum and some suboptimum phase estimators for the following cases: (1) known constant frequency and unknown constant phase with an a priori distribution; (2) unknown constant frequency and phase with a joint a priori distribution; (3) frequency a parameterized function of time with a joint a priori distribution on parameters and phase; (4) frequency a gaussian random process. (Part I introduces the general problem and treats case 1).

I. Introduction

Many of the algorithms, such as the phase-locked loop, currently used for phase estimation were originated against the background of analog technology some 25 years ago. While most of the algorithms have proved to be readily implementable in terms of current digital technology, this does not necessarily mean these are the most desirable algorithms with current (or near future) technology. In some cases, the structure of the optimal phase estimator is not complicated and it is worth considering against the background of new technology, whether some of the optimal or related suboptimal structures may be more desirable than some of the older algorithms. The purpose of this article is to look at some of the optimal and suboptimal structures.

II. Problem Statement

Following the approach of Ref. 1, we choose for the modulator function

$$y\,(f_c t + x_0(t))$$

where $y\,(\cdot)$ is periodic with unit period, f_c is the nominal frequency of the periodic waveform and $x_0(t)$ is the phase to be estimated on the basis of the received signal

$$z(t) = y\,(f_c t + x_0(t)) + n(t)$$

where $n(t)$ is a gaussian random process with mean 0 and known autocorrelation. In the case where the frequency is constant and known equal to f_c, then $x_0(t) = x_0$ is a constant to be estimated by a certain functional on $z(t)$ (random variable). In the case where the frequency is constant and unknown, then $x(t) = f_d t + x_0$ is a ramp function with parameters f_d and x_0 to be estimated in the form of certain functionals on $z(t)$. In the case where $x_0(t)$ is a gaussian random process with known autocorrelation function, then the estimator for $x(t)$ is a certain operator on $z(t)$, which is of course a random process.

Let us assume the channel for $z(t)$ is bandlimited with bandlimit $F/2$ and that the gaussian noise $n(t)$ is white over this bandwidth. Then to proceed with the analysis we shall deal with all waveforms in terms of their unique sample representations at the sampling rate F. Thus,

$$\left.\begin{aligned} t_i &= i/F \\ z_i &= z(i/F) \\ n_i &= n(i/F) \\ y_i &= y(i/F) \\ x_i &= x(i/F) \end{aligned}\right\} \tag{1}$$

where the index i runs over the integers. For $x_0(t)$, a gaussian random process, the multivariant a priori probability density for the phase sequence $\{x_i\}$, $A_x \leqslant i \leqslant B_x$ is given by

$$f_x(\{x_i\}) = (2\pi)^{-\frac{B_x - A_x + 1}{2}} |R_{xij}|^{-1/2}$$

$$\exp\left\{-1/2 \sum_{i=A_x}^{B_x} \sum_{j=A_x}^{B_x} R_{xij}^{-1} x_i x_j\right\} \tag{2}$$

where $R_{xij} = R_{xx}(i/F, j/F)$ is the covariance matrix, $|R_{xij}|$ is its determinant and R_{xij}^{-1} is its inverse. In the case of unknown constant frequency and phase, we drop the gaussian assumption and the joint a priori probability density for f_0 and x_0 is given by

$$f_{fx}(f_0, x_0) = f_f(f_0) \cdot f_x(x_0) \tag{3}$$

where we have assumed f_0 and x_0 are independent random variables for the usual application and $f_f(\cdot)$ and $f_x(\cdot)$ are arbitrary. In the case of known constant frequency f_c and unknown phase x_0, the a priori probability density

$$f_x(x_0) \tag{4}$$

may also be chosen arbitrarily.

Similarly to Eq. (2) the probability density function for the noise sequence $\{n_i\}$ is given by

$$f_n(\{n_i\}) = (2\pi)^{-\frac{B_z - A_z + 1}{2}} |R_{nij}|^{-1/2}$$

$$\exp\left\{-1/2 \sum_{i=A_z}^{B_z} \sum_{j=A_z}^{B_z} R_{nij}^{-1} n_i n_j\right\} \tag{5}$$

Because of the assumptions regarding noise spectrum and sampling rate, the members of the noise sequence are uncorrelated.

This gives

$$R_{nij} = \sigma_n^2 \, \delta_{ij} \tag{6}$$

$$R_{nij}^{-1} = \frac{1}{\sigma_n^2} \delta_{ij} \tag{7}$$

$$|R_{nij}| = \sigma_n^{2(B_z - A_z + 1)} \tag{8}$$

where δ_{ij} is the Kronecker delta and we have used the assumption that $n(t)$ is wide sense stationary. Substituting Eqs. (6), (7), and (8) into Eq. (5) gives

$$f_n(\{n_i\}) = (2\pi)^{-\frac{B_z - A_z + 1}{2}} \sigma^{-(B_z - A_z + 1)}$$

$$\exp\left\{-\frac{1}{2\sigma_n^2} \sum_{i=A_z}^{B_z} n_i^2\right\} \tag{9}$$

Under the assumption that $\{x_i\}$ and $\{n_i\}$ are independent random processes

$$f_{z|x}(\{z_i\} \mid \{x_i\}) = f_n(\{z_i - y(f_c t + x_i)\}) \tag{10}$$

and

$$f_{zx}(\{z_i\}, \{x_i\}) = f_n(\{z_i - y(f_c t + x_i)\})\, f_x(\{x_i\}) \tag{11}$$

Then

$$f_{x|z}(\{x_i\} \mid \{z_i\}) = \frac{f_n(\{z_i - y(f_c t + x_i)\}) f_x\{x_i\}}{f_z(\{z_i\})} \tag{12}$$

Now the denominator of Eq. (12) is not a function of $\{x_i\}$, it is merely a factor that normalizes the joint density of Eq. (11) with respect to $\{x_i\}$. Thus, when estimating $\{x_i\}$ on the basis of a received $\{z_i\}$, we may treat $f_z(\{z_i\})$ as a constant.

Substituting Eqs. (2) and (9) in Eq. (12) gives

$$f_{x|z}(\{x_i\}|\{z_i\}) = C_1(\{z_i\})$$

$$\exp(-1/2)\left[\sum_{i=A_x}^{B_x}\sum_{j=A_x}^{B_x} R_{xij}^{-1} x_i x_j\right.$$

$$\left.+\frac{1}{\sigma_n^2}\sum_{k=A_z}^{B_z}(z_k - y(f_c t_k + x_k))^2\right] \tag{13a}$$

where $C_1(\{z_i\})$ includes the constant coefficients from Eqs. (2) and (9). For the case of constant unknown frequency and phase we use Eq. (3) instead of Eq. (2) which gives, in place of Eq. (13a),

$$f_{fx|z}(f, x \mid \{z_i\}) = C_2(\{z_i\})$$

$$\exp\left[-\frac{1}{2\sigma_n^2}\sum_{k=A_z}^{B_z}(z_k - y(ft_k + x))^2\right.$$

$$\left.+\ln f_f(f) + \ln f_x(x)\right] \tag{13b}$$

and, for the case of constant known frequency and unknown phase,

$$f_{x|z}(x|\{z_i\}) = C_2(\{z_i\})$$

$$\exp\left[\frac{-1}{2\sigma_n^2}\sum_{k=A_z}^{B_z}(z_k - y(f_c t_k + x))^2\right.$$

$$\left.+\ln f_x(x)\right] \tag{13c}$$

The conditional probability density functions (13) contain all needed information for estimation of the phase (and/or frequency). The maximum a posteriori probability (MAP) estimator is simply the mode of the distribution (13), given the observed $\{z_i\}$. The minimum mean square error (MMSE) estimator is the mean of the distribution and the minimum mean absolute error (MMAE) estimator is the median of the distribution, etc. It is possible for a distribution to fail to have a unique mode or median.

III. Estimation of Unknown Phase with Known Constant Frequency

As the sampling rate F and noise bandwidth $F/2$ increase, while holding constant the noise spectral density $S_{nn} = \sigma_n^2/F$, the initial time $t_1 = A_z/F$ and the final time $t_2 = B_z/F$, the summation in Eq. (13c) becomes well approximated by an integral and we can write

$$f_{x|z}(x|[z(\cdot)]) = C_3[z(t)]$$

$$\exp\left[\frac{-1}{2S_{nn}}\int_{t_1}^{t_2}(z(\tau)\right.$$

$$\left.- y(f_c\tau + x))^2 d\tau + \ln f_x(x)\right] \tag{14}$$

where the constant (with respect to x) coefficient C_3 is now a functional on $z(t)$. (As F becomes large, the coefficient for $f_{x,z}(\cdot,\cdot)$ becomes small, but the coefficient for $f_{x|z}(\cdot|\cdot)$ does not.) If the argument of the exponential function in Eq. (14) is an even function of x about its maximum value, then the value of x at the maximum is not only the MAP estimator $\hat{x}_0$, but also the MMSE and MMAE estimator. We will pursue the MAP estimator and observe when it is also the optimum estimator under other criteria.

From Eq. (14) we see that the MAP estimator, if it exists, is the value of x that maximizes

$$-\frac{1}{N_0}\int_{t_1}^{t_2} (z(\tau) - y(f_c\tau + x))^2 \, d\tau + \ln f_x(x) \tag{15}$$

where N_0 is the one-sided noise spectral density. Expanding the integrand gives

$$-\frac{1}{N_0}\int_{t_1}^{t_2} z^2(\tau)\, d\tau + \frac{2}{N_0}\int_{t_1}^{t_2} z(\tau)\, y(f_c\tau + x)\, d\tau$$

$$-\frac{1}{N_0}\int_{t_1}^{t_2} y^2(f_c t + x)\, d\tau + \ln f_x(x) \tag{16}$$

The first integral is not a function of x and the third integral is not a function of x if $t_2 - t_1 = n/f_c$, n a positive integer, which we assume henceforth. Thus the MAP estimator $\hat{x}_0$ is the value of x that maximizes

$$\frac{2}{N_0}\int_{t_1}^{t_2} z(\tau) y(f_c\tau + x)\, d\tau + \ln f_x(x) \tag{17}$$

Real functions such as $z(t)$ and $y(f_c t + x)$ on the interval $t_1 \leqslant t \leqslant t_2$ are elements of a signal space (essentially a Hilbert space) with the usual inner product

$$\langle \vec{x}, \vec{y} \rangle = \int_{t_1}^{t_2} x(t)\, y(t)\, dt \tag{18}$$

and norm

$$\|\vec{x}\| = \sqrt{\langle \vec{x}, \vec{x} \rangle} \tag{19}$$

In terms of the vector space notation, Eq. (15) can be written

$$-\|\vec{z} - \vec{y}(x)\|^2 + N_0 \ln f_x(x) \tag{20}$$

and Eq. (17) can be written

$$2\langle \vec{z}, \vec{y}(x) \rangle + N_0 \ln f_x(x) \tag{21}$$

where $\|\vec{z}\|$ and $\|\vec{y}(x)\|$ are not functions of x. If x_0 has a uniform a priori distribution

$$f_x(x_0) = \begin{cases} 1, & -1/2 \leqslant x_0 \leqslant 1/2 \\ 0, & \text{elsewhere} \end{cases} \tag{22}$$

as is usually the case for an initial estimate, then we see from Eq. (20) that $\hat{x}_0$ should be chosen to minimize the distance between $\vec{z}$ and $\vec{y}(\hat{x}_0)$ or, from Eq. (21), to minimize the angle between $\vec{z}$ and $\vec{y}(\hat{x}_0)$ (maximize the cosine of the angle where

$$\left.\cos\theta = \frac{\langle \vec{z}, \vec{y}(\hat{x}_0) \rangle}{\|\vec{z}\| \cdot \|\vec{y}\|}\right) \tag{23}$$

If $y(\cdot)$ is differentiable for all values of its argument then a necessary condition for maximizing

$$\langle \vec{z}, \vec{y}(x) \rangle = \int_{t_1}^{t_2} z(\tau)\, y\,(f_c\tau + x)\, d\tau \tag{24}$$

is

$$\langle \vec{z}, \vec{y}'(x) \rangle = \int_{t_1}^{t_2} z(\tau)\, y'(f_c\tau + x)\, d\tau = 0 \tag{25}$$

Schemes that maximize Eq. (24) are often called direct estimators and those that solve Eq. (25) are often called indirect estimators. Indirect estimation must be used with caution. In the first place, the integrand of Eq. (25) must contain the derivative of $y(\cdot)$ and not merely $y(\cdot)$ shifted one quarter period, as in the special case of a sinusoid. Also Eq. (25) is a necessary condition and may well be satisfied by values of x that do not maximize Eq. (24). For example, Eq. (24) may have a number of relative maxima only one of which is the absolute maximum. In this case, Eq. (25) has multiple solutions.

If $|y(\cdot)|$ is a bounded variation function, as we shall assume, then the tip of the vector $\vec{y}(x)$ describes a continuous closed curve on the surface of a hypersphere of radius $\|\vec{y}\|$ about the origin in the signal space as x traverses its unit interval from $-1/2$ to $+1/2$. This curve may not have a derivative everywhere and it spans a linear subspace L_y that is generally not finite dimensional. The MAP phase direct estimation problem is now visualized by choosing a point $\vec{y}(x_0)$ on this curve (the phase to be estimated), moving away a distance in some direction to $\vec{z}$, determined by the additive noise vector[1] $\vec{n}$

$$\vec{z} = \vec{y}(x_0) + \vec{n} \tag{26}$$

and then seeking the point $\hat{x}_0$ on the curve $\vec{y}(x)$, parameterized by x, which is closest to $\vec{z}$ or its orthogonal projection onto L_y. In equation form this amounts to choosing $\hat{x}_0$ to minimize

$$\|\vec{y}(x_0) + \vec{n} - \vec{y}(\hat{x}_0)\|$$

(see Eq. (20) with $f_x(\cdot)$ constant). This is the same as choosing the point $\vec{y}(\hat{x}_0)$ on the curve to minimize the angle between $\vec{z}$ (or its orthogonal projection onto L_y) and $\vec{y}(\hat{x}_0)$ (see Eq. (21)). Since the path described by $\vec{y}(x)$ on the surface of the hypersphere can be rather arbitrary, depending on the choice of $y(\cdot)$, it can be appreciated that the indirect method of estimation that looks for points on the curve for which $\|\vec{z} - \vec{y}(\hat{x}_0)\|$ or $\langle\vec{z}, \vec{y}(\hat{x}_0)\rangle$ is stationary must be used with caution. The point $\hat{x}_0$ that maximizes $\langle\vec{z}, \vec{y}(x)\rangle$ may not even be stationary and there may be stationary points that do not give the maximum value of $\langle\vec{z}, \vec{y}(x)\rangle$. The vector $\vec{y}'(x)$, if it exists, is tangent to the curve and the solution of Eq. (25) involves finding points on the curve where its tangent is orthogonal to z (or its orthogonal projection onto L_y).

The indirect method tends to be computationally more desirable than the direct method (driving something to zero tends to be simpler than driving something to its global maximum). For this reason, many phase estimation devices are based on or inspired by the indirect method. A well-known example inspired by the indirect method is the phased-locked loop.

As we shall see in Section V, when $y(\cdot)$ is sinusoidal all of the perils of the indirect method disappear and other good things happen. But for as simple a waveform as the square wave there are problems as discussed in Section VI.

[1]We need only consider the orthogonal projection of $\vec{n}$ onto the linear subspace L_y spanned $\vec{y}(x)$.

IV. Discrete Phase Estimation

In case there are only a finite number of possible phases, as in the case of subcarrier or symbol synchronization in a fully synchronous system, the indirect method cannot be used if advantage is to be taken of the a priori information. The possible phases $x_{01}, x_{02}, \cdots, x_{0n}$ are used to obtain n values of Eq. (17), with $f_x(\cdot)$ replaced by the discrete probabilities, and the largest value of Eq. (17) identified (not more than two values of Eq. (17) need be stored at any time). This use of the direct method does not require that $y(\cdot)$ be differentiable, etc.

V. Estimation Of The Phase Of A Sinusoid

If $y(\eta) = A_c \cos 2\pi\eta$ then the curve described by $\vec{y}(x)$ in signal space is a circle of radius $A_c\sqrt{T/2}$ about the origin and lying in a certain two-dimensional linear subspace L_y (a plane through the origin). The orientation of the plane of the circle with respect to a fixed basis depends on the frequency f_c. The direct method of estimation amounts to finding the point $\vec{y}(\hat{x}_0)$ on the circle closest to the received signal $\vec{z} = \vec{y}(x_0) + \vec{n}$ or its orthogonal projection onto L_y and the indirect method amounts to finding a stationary point where the tangent of the circle is orthogonal to $\vec{z}$ or its orthogonal projection onto L_y. If $\vec{z}$ is not orthogonal to the plane of the circle (a singular condition) then there are always exactly two stationary points and the one corresponding to minimum $\|\vec{z} - \vec{y}(\hat{x}_0)\|$ is easily selected. The fact that $\vec{y}(x)$ lies in a two-dimensional linear subspace gives the further simplification that $\|\vec{z} - \vec{y}(x)\|$ and $\langle\vec{z}, \vec{y}(x)\rangle$ can be evaluated in terms of two scalar functions of $\vec{z}$ (functionals of $z(t)$) and two trigonometric functions of x. In fact Eq. (25) can be solved explicitly for x_0 in terms of the two scalar functions of $\vec{z}$ (the coordinates of $\vec{z}$ with respect to a basis). Also, due to the symmetry of the circular path of $\vec{y}(x)$, $\|\vec{z} - \vec{y}(x)\|$ is an even function about its minimum for any $\vec{z}$ and thus, by Eq. (14),

$$f_{x|z}(x|\vec{z}) = C_3(\vec{z}) \exp\left[\frac{-1}{2S_{nn}} \|\vec{z} - \vec{y}(x)\|^2 + \ln f(x)\right] \tag{27}$$

is an even function about its maximum provided the a priori distribution $f(x)$ is uniform. In this case, the MAP estimator is also the MMSE, MMAE, etc. estimator.

Let us return to Eq. (17) to obtain the indirect estimator relation by differentiating

$$\frac{2}{N_0}\int_{t_1}^{t_2} z(\tau)A_c \cos 2\pi(f_c\tau + x)\,d\tau + \ln f_x(x) \tag{28}$$

to obtain

$$\frac{2 \cdot 2\pi A_c}{N_0} \int_{t_1}^{t_2} z(\tau) \sin 2\pi (f_c \tau + \hat{x}_0)\, d\tau = \frac{f_x'(\hat{x}_0)}{f(\hat{x}_0)} \quad (29)$$

Define

$$I[z(\cdot)] = \int_{t_1}^{t_2} z(\tau) \cos 2\pi f_c \tau\, d\tau = \frac{\langle \vec{z}, \vec{y}(0) \rangle}{\sqrt{2}\, \|\vec{y}\|} \quad (30)$$

$$Q[z(\cdot)] = \int_{t_1}^{t_2} z(\tau) \sin 2\pi f_c \tau\, d\tau = -\frac{\langle \vec{z}, \vec{y}'(0) \rangle}{\sqrt{2}\, \|\vec{y}'\|} \quad (31)$$

Then Eq. (29) may be written

$$Q \cos 2\pi \hat{x}_0 + I \sin 2\pi \hat{x}_0 = \frac{N_0}{4\pi A_c} \frac{f_x'(\hat{x}_0)}{f_x(\hat{x}_0)} \quad (32)$$

If the signal-to-noise ratio is very large or if the a priori distribution is uniform the right member of Eq. (32) vanishes and we have the familiar result

$$2\pi \hat{x}_0 = -\arctan \frac{Q}{I} \quad (33)$$

With current digital technology Eq. (33) can be easily mechanized and x_0 can be obtained on a four-quadrant basis.

It is interesting to consider the functional form of $f_x'(x)/f_x(x)$ for several a priori distributions. In the case of the gaussian distribution

$$\frac{f_x'(x)}{f_x(x)} = -\frac{x - m}{\sigma^2} \quad (34)$$

where m is the mean and σ^2 the variance. The gaussian is the only distribution giving a linear result. Substituting Eq. (34) into the right member of Eq. (29) gives something like a phase-locked loop except that $\hat{x}_0$ is constant over the interval $t_1 \leqslant t \leqslant t_2$ instead of a function of time.

The a posteriori density function for the estimator of Eq. (33) is given by (Ref. 2)[2] :

$$f_{x|\hat{x},r}(x_0|\hat{x}_0, r) = C\, e^{\alpha r \cos 2\pi (x_0 - \hat{x}_0)} \quad \text{(footnote 3)} \quad (35)$$

where C is the normalizing constant and

$$\alpha^2 = \frac{A_c^2 T}{N_0} \quad (36)$$

$$r^2 = 4 \frac{I^2 + Q^2}{N_0 T} \quad (37)$$

where $T = t_2 - t_1$.

The distribution (35) can also be viewed as the a priori distribution for a second estimate $\hat{x}_{02}$ conditioned on the first estimate $\hat{x}_{01}$ $(= \hat{x}_0)$. We can then write

$$f_{x_2|x_1,r_1}(\hat{x}_{02}|\hat{x}_{01}, r_1) = C\, e^{\alpha_1 r_1 \cos 2\pi (\hat{x}_{02} - \hat{x}_{01})} \quad (38)$$

where we assume I_2 and Q_2 are independent of I_1 and Q_1 as in the case where the time intervals for the successive estimations are nonoverlapping. From Eq. (38),

$$\frac{f'_{x_2|x_1,r_1}(\hat{x}_{02}|\hat{x}_{01}, r_1)}{f_{x_2|x_1,r_1}(\hat{x}_{02}|\hat{x}_{01}, r_1)} = -2\,\alpha_1 r_1 \sin 2\pi (\hat{x}_{02} - \hat{x}_{01}) \quad (39)$$

[2]The distribution Eq. (35) results from averaging Eq. (14) over all $z(t)$ subject to the condition of the observable r. If the condition r is removed then

$$f_{x|\hat{x}}(x_0|\hat{x}_0) = e^{-\frac{\alpha^2}{2}} + \sqrt{2\pi}\, \alpha \cos 2\pi(\hat{x}_0 - x_0)\, \Phi\left(\alpha \cos 2\pi(\hat{x}_0 - x_0)\right) e^{-\frac{\alpha^2}{2} \sin^2 2\pi(\hat{x}_0 - x_0)}$$

which is less desirable.

[3]$C\, e^{\alpha r \cos 2\pi\epsilon}$ is simply the probability density function of the estimation error ϵ, conditioned on r.

which can be substituted in the right member of Eq. (32) to give

$$Q_2 \cos 2\pi\hat{x}_{02} + I_2 \sin 2\pi\hat{x}_{02} = -\frac{N_0 \alpha_1 r_1}{2A_c} \sin 2\pi(\hat{x}_{02} - \hat{x}_{01}) \tag{40}$$

Expanding the sine term in the right member and solving for x_{02} gives

$$\tan 2\pi\hat{x}_{02} = -\frac{Q_2 + \dfrac{N_0\alpha_1 r_1}{2A_c}\sin 2\pi\hat{x}_{01}}{I_2 + \dfrac{N_0\alpha_1 r_1}{2A_c}\cos 2\pi\hat{x}_{01}} \tag{41}$$

This is actually a bayesian recursive estimator based on the new observables Q_2 and I_2 and the previous estimate $\hat{x}_{01}$ and the previous r_1. A little calculation will show that

$$\frac{N_0\alpha_1 r_1}{2A_c} \sin 2\pi\hat{x}_{01} = Q_1, \; \frac{N_0\alpha_1 r_1}{2A_c} \cos 2\pi\hat{x}_{01} = I_1 \tag{42}$$

and so Eq. (41) can be written

$$\tan 2\pi\hat{x}_{02} = -\frac{Q_2 + Q_1}{I_2 + I_1} \tag{43}$$

which is as it should be since the optimum recursive estimator should give the same result as a one-time optimum estimator using all of the observables. Clearly the error distribution for $\hat{x}_{02}$ has the same form [Eq. (35)] as that for $\hat{x}_{01}$ and thus the form of Eq. (35) is a reproducing distribution for this recursive baysian estimator. At each step of the estimator, $f_{x|z}(x_{0n} | z(\cdot))$ is an even function of x_{0n} about its maximum [see Eq. (14)] and so the MAP recursive estimator is also MMSE, MMAE, etc.

The mechanization of an estimator suggested by Eq. (43) and Eq. (35) seems straightforward. Successive pairs Q_i, I_i for successive equal increments of time of duration T are stored in a shift register and the contents summed to give

$$Q = \sum_{i=1}^{n} Q_i, \quad I = \sum_{i=1}^{n} I_i \tag{44}$$

where n is increasing with time. The phase estimate at any time is given by

$$\tan 2\pi\hat{x}_0 = -\frac{Q}{I} \tag{45}$$

and the error of the estimate is described by Eq. (35) where

$$\alpha r = \frac{2A_c}{N_0} \sqrt{(I^2 + Q^2)} \tag{46}$$

The mean square error of the estimate (or any other moment) is an easily calculated function of αr. Thus if a specified value of the MSE is designated as the "operating threshold," then when the generally increasing αr reaches the corresponding value, the two shift registers can start dumping so that Eq. (44) is replaced by

$$Q = \sum_{n-N}^{n} Q_i, \quad I = \sum_{n-N}^{n} I_i \tag{47}$$

to give a running or aperture sum of the Q_i, I_i. The calculation of αr from Eqs. (46) and (47) is continued for each value of n and, if it should decrease significantly due to the run of values of Q_i^2 and I_i^2 (r is a random variable), then the value of N in Eq. (47) can be automatically increased.

This has the advantage of following slow changes in x_0 with an aperture filter of length NT with N controlled to maintain the quality of the estimate at a prescribed level. This is something like a phase-locked loop of variable bandwidth with a guaranteed limit on noise MSE. However, unlike the phase-locked loop, there is not a phase initial condition to affect acquisition time.

The aperture averaging scheme is a suboptimal way of handling the estimation of a nonconstant phase. If a variable phase $x(t)$ can be modelled by a parameterized random process such as

$$x(t) = f_d t + x_0$$

with f_d and x_0 random variables, or by a more general one, then an optimum estimator can be formulated as discussed in Part II.

VI. Estimation of the Phase of Nonsinusoids

Returning to Eqs. (20) and (21), we see that to evaluate $\|\vec{z}_y - \vec{y}(x)\|$ or $\langle \vec{z}_y, \vec{y}(x) \rangle$ as functions of $\vec{z}_y$ we must have the orthogonal projection $\vec{z}_y$ of $\vec{z}$ onto the linear subspace L_y spanned by $\vec{y}(x)$, $-1/2 \leqslant x \leqslant 1/2$ expressed in terms of coordinates $z_1, z_2, \cdots$ of $\vec{z}$ with respect to some orthonormal basis $\vec{v}_1, \vec{v}_2, \cdots$ for L_y. For the sinusoid considered in the previous section, L_y is two dimensional, and orthonormal basis vectors $\vec{v}_1$ and $\vec{v}_2$ are given by [see Eqs. (30) and (31)]:

$$v_1(t) = \sqrt{\frac{2}{T}} \cos 2\pi f_c t \tag{48}$$

$$v_2(t) = \sqrt{\frac{2}{T}} \sin 2\pi f_c t \tag{49}$$

and the coordinates z_1 and z_2 of $\vec{z}$ are given by

$$z_1 = \langle \vec{z}, \vec{v}_1 \rangle = \sqrt{\frac{2}{T}}\, I\,[z(\cdot)] \tag{50}$$

$$z_2 = \langle \vec{z}, \vec{v}_2 \rangle = \sqrt{\frac{2}{T}}\, Q\,[z(\cdot)] \tag{51}$$

These coordinates of $\vec{z}$ are a minimum set of observables required to determine $\hat{x}_0$ as in Eq. (33) or Eq. (41). When L_y has $n \geqslant 3$ dimensions, the explicit solution for $\hat{x}_0$ in terms of the n observables (coordinates of $\vec{z}$) becomes more difficult.

A convenient orthonormal basis for a general periodic $y(\eta)$ is the fourier basis $\sqrt{2/T}$ cos $2\pi j\eta$, $\sqrt{2/T}$ sin $2\pi j\eta$, $j = 1, 2, \cdots \infty$. Clearly any $y(\cdot)$ with a discontinuity or a discontinuity in any derivative requires a basis of infinite dimension. In such a case, the minimum set of observables required to determine $\hat{x}_0$ is, in effect, the complete function $z(t)$ from t_1 to t_2; that is, the infinite set of coordinates of $\vec{z}$ with respect to the basis above. In the case of large noise where the magnitude of the orthogonal projection of $\vec{n}$ onto L_y is of the same order as $\|\vec{y}\|$, other periodic waveforms $y(\cdot)$ cannot be much better than the sinusoid for phase estimation. This is because the difference between the maximum and minimum values of $\|\vec{z}_y - \vec{y}(x)\|$ tend to be about as large for the sinusoid as for any other $y(\cdot)$. The sinusoid describes a great circle on the surface of the hypersphere of radius $\|\vec{y}\|$ and the path of any other $\vec{y}(x)$ is also confined to the surface of the same hypersphere.

However, for the case of small noise where $\vec{n}_y$ usually does not take $\vec{z}$ far from the path $\vec{y}(x)$ on the hypersphere and the error due to the noise is approximately the component of $\vec{n}$ along $\vec{y}'(x_0)$, the error in the estimate can be reduced simply by increasing the length of the closed path described by $\vec{y}(x)$ so that a given noise displacement along $\vec{y}'(x_0)$ corresponds to a smaller increment of x. This can only be done by increasing the dimension of L_y and of the resulting one-less dimension of the surface of the hypersphere on which the path described by $\vec{y}(x)$ lies. If s is the arc length along the path, then we want to increase

$$\frac{ds}{dx} = \|\vec{y}'(x)\| \tag{52}$$

in order to reduce the error of phase estimation for the small noise case. For the sinusoidal case of the previous section,

$$\frac{ds}{dx} = 2\pi A_c \sqrt{\frac{T}{2}} \tag{53}$$

and the variance of the zero mean noise component along s is just $S_{nn} = N_0/2$. Thus

$$\sigma^2(\hat{x}_0) = \frac{N_0/2}{\left(\dfrac{ds}{dx}\right)^2} = \frac{1}{4\pi^2} \cdot \frac{N_0}{A_c^2 T} \tag{54}$$

a well known result.

For the case of

$$y(\eta) = \sum_{j=1}^{n} (a_j \cos 2\pi j\eta + b_j \sin 2\pi j\eta) \tag{55}$$

where

$$(1/2) \sum_{j=1}^{n} (a_j^2 + b_j^2) = p_y \text{ is fixed}$$

we have, by Eq. (52),

$$\left(\frac{ds}{dx}\right)^2 = 2\pi^2 T \sum_{j=1}^{n} j^2 (a_j^2 + b_j^2) \tag{56}$$

As a simple example of a nonsinusoidal $y(\cdot)$ let us consider

$$y(\eta) = \sqrt{0.9}\,A_c \left(\cos 2\pi\eta + \frac{1}{3}\cos 6\pi\eta\right) \tag{57}$$

which has the same power as $A_c \cos 2\pi\eta$ and which is the fundamental and next higher harmonic of a square wave. Here

$$\left(\frac{ds}{dx}\right)^2 = 0.9 \times 4\pi^2 A_c^2 T \tag{58}$$

and

$$\sigma^2(\hat{x}_0) = \frac{1}{0.9 \times 8\pi^2} \cdot \frac{N_0}{A_c^2 T} \tag{59}$$

for small noise. This is a 2.6-db reduction of phase error with respect to the sinusoid result of Eq. (54). We next consider phase estimation for this example in the presence of large noise. The L_y for Eq. (57) has four dimensions and $\vec{y}(x)$ traces out a smooth (all derivatives exist) closed path on the three-dimensional surface of a four-dimensional hypersphere for $-1/2 \leqslant x \leqslant 1/2$. The path is rather more interesting than the circle considered in the previous section. For a typical large noise $\vec{z}$ the norm $\|\vec{z}_y - \vec{y}(x)\|$ to be minimized as a function of x has six stationary points – three maxima and three minima. If we define

$$I_1 = \int_{t_1}^{t_2} z(\tau) \cos 2\pi f_c \tau \, d\tau \tag{60a}$$

$$Q_1 = \int_{t_1}^{t_2} z(\tau) \sin 2\pi f_c \tau \, d\tau \tag{60b}$$

$$I_3 = \int_{t_1}^{t_2} z(\tau) \cos 6\pi f_c \tau \, d\tau \tag{60c}$$

$$Q_3 = \int_{t_1}^{t_2} z(\tau) \sin 6\pi f_c \tau \, d\tau \tag{60d}$$

then

$$\frac{\langle \vec{z}_y, \vec{y}(x) \rangle}{\sqrt{0.9}\,A_c} = I_1 \cos 2\pi x - Q_1 \sin 2\pi x + \frac{I_3}{3} \cos 6\pi x - \frac{Q_3}{3} \sin 6\pi x \tag{61}$$

and

$$\frac{\langle \vec{z}_y, \vec{y}'(x) \rangle}{\sqrt{0.9}\,A_c} = -2\pi \left[Q_1 \cos 2\pi x + I_1 \sin 2\pi x + Q_3 \cos 6\pi x + I_3 \sin 6\pi x \right] \tag{62}$$

For the case $I_1 = Q_1 = I_3 = Q_3 = 1$ the indirect estimator $\langle \vec{z}_y, \vec{y}'(x) \rangle = 0$ has six solutions at approximately $x = -0.31$, -0.25, $-.06$, 0.19, 0.25, 0.44 at which are located respectively a max, min, max, min, max, min of $\langle \vec{z}_y, \vec{y}(x) \rangle$. The respective values of $\langle \vec{z}_y, \vec{y}(x) \rangle / \sqrt{0.9}\,A_c$ are 0.72, 0.67, 1.74, -0.72, -0.67, -1.74. Of the six values of x provided by the indirect estimator only one, $\hat{x}_0 = -0.06$ is the MAP estimate of x_0. In this case $\langle \vec{z}_y, \vec{y}(x) \rangle$ is not an even function with respect to $x = \hat{x}_0$ and therefore Eq. (14) (with $f_x(\cdot)$ uniform) is not an even function with respect to its maximum at $x = \hat{x}_0$. Thus it is not clear whether the MAP estimator is MMSE, MMAE, etc. As we have seen, it is not difficult to find $\sigma^2(\hat{x}_0)$ for small noise (in this case $\hat{x}_0$ is approaching the MMSE estimator), but the probability distribution for $\hat{x}_0$ does not appear to be known for the general case.

For small noise, the indirect MAP estimator $\langle \vec{z}_y, \vec{y}'(x) \rangle = 0$ may have six solutions, but if so, there will be two closely spaced pairs that are easily distinguished and rejected while the other two are as in the case of a sinusoid. The small noise performance of the MAP estimator above in Eq. (59) is considerably better (7-db) than that of the suboptimal estimator $\langle \vec{z}, \vec{y}(x - 1/4) \rangle = 0$, which is sometimes used.

We turn next to estimation of the phase of a square wave

$$y(\eta) = \mathrm{Cos}\, 2\pi\eta \quad \text{(footnote 4)} \tag{63}$$

[4]The Notation Cos and Sin refers to unit-amplitude square waves in phase with cos and sin.

Here L_y is infinite dimensional, and the closed path on the hypersphere surface described by $\vec{y}(x)$ is the sum of circular motions in mutually orthogonal planes in accordance with the fourier series expansion of Eq. (63). In this case, the maximum of $\langle\vec{z},\vec{y}(x)\rangle$ [see Eq. (21)] is generally not stationary and an indirect MAP estimator is not appropriate. The direct estimator for a uniform a priori distribution on x_0 seeks the value of x that maximizes

$$\langle\vec{z}_y,\vec{y}(x)\rangle = \int_{t_1}^{t_2} z(\tau)\ \mathrm{Cos}\ 2\pi(f_c\tau + x)\ d\tau$$

$$= \frac{4}{\pi}\sum_{j=1}^{\infty}\frac{1}{2j-1}\int_{t_1}^{t_2} z(\tau)\cos 2\pi(2j-1)$$

$$\times (f_c\tau + x)\ d\tau \tag{64}$$

$$= \frac{4}{\pi}\sum_{j=1}^{\infty}\frac{I_{2j-1}}{2j-1}\cos 2\pi(2j-1)x$$

$$- \frac{Q_{2j-1}}{2j-1}\sin 2\pi(2j-1)x$$

where

$$I_k = \int_{t_1}^{t_2} z(\tau)\cos 2\pi k f_c\tau\ d\tau \tag{65a}$$

$$Q_k = \int_{t_1}^{t_2} z(\tau)\sin 2\pi k f_c\tau\ d\tau \tag{65b}$$

much as in Eq. (60), where the I_k, Q_k are the infinite (or very large) set of observables needed. Since

$$z(t) = A_c\ \mathrm{Cos}\ 2\pi(f_c t + x_0) + n(t) \tag{66}$$

I_k and Q_k are independent gaussian random variables with means

$$E[I_k] = \frac{2A_cT}{\pi k}\cos 2\pi k x_0 \tag{67a}$$

$$E[Q_k] = -\frac{2A_cT}{\pi k}\sin 2\pi k x_0 \tag{67b}$$

and variances

$$\sigma^2[I_k] = \sigma^2[Q_k] = \frac{N_0T}{4} \tag{68}$$

where $T = t_2 - t_1$. Another way to look at this estimation problem is to use Eq. (66) to write

$$\langle\vec{z}_y,\vec{y}(x)\rangle = A_c\int_{t_1}^{t_2}\mathrm{Cos}\ 2\pi(f_c\tau + x_0)\ \mathrm{Cos}\ 2\pi(f_c\tau + x)\ d\tau$$

$$+ \int_{t_1}^{t_2} n(\tau)\ \mathrm{Cos}\ 2\pi(f_c\tau + x)\ d\tau \tag{69}$$

The first integral is an even periodic triangular function of $x - x_0$ of period unity, peak values $\pm A_cT$, and slopes $\pm 4A_cT$. The last integral is a periodic random process $N_z(x)$ that, when added to the triangular function, is the function of x to be maximized for the MAP estimator. Thus

$$\langle\vec{z}_y,\vec{y}(x)\rangle = A_cT\ \mathrm{Cos}\ 2\pi(x - x_0) + N_z(x)\quad \text{(footnote 5)} \tag{70}$$

Now

$$R_{NN}(x_1,x_2) = E\left[\int_{t_1}^{t_2}\int_{t_1}^{t_2} n(\tau_1)\,n(\tau_2)\ \mathrm{Cos}\ 2\pi(f_c\tau_1 + x_1)\right.$$

$$\left.\times\ \mathrm{Cos}\ 2\pi(f_c\tau_2 + x_2)\ d\tau_2 d\tau_1\right]$$

$$= \frac{N_0}{2}\int_{t_1}^{t_2}\int_{t_1}^{t_2}\delta(\tau_1 - \tau_2)\ \mathrm{Cos}\ 2\pi(f_c\tau_1 + x_1)$$

[5] The notation Cos and Sin refers to unit-amplitude triangular waves in phase with cos and sin.

$$\times \operatorname{Cos} 2\pi(f_c\tau_2 + x_2)\, d\tau_2 d\tau_1$$

$$= \frac{N_0}{2} \int_{t_1}^{t_2} \operatorname{Cos} 2\pi(f_c\tau_1 + x_1)$$

$$\times \operatorname{Cos} 2\pi(f_c\tau_1 + x_2)\, d\tau_1$$

$$= \frac{N_0 T}{2} \operatorname{Cos} 2\pi(x_1 - x_2) \tag{71}$$

Thus the autocorrelation of $N_z(x)$ is a periodic triangular function of unit period. This means the power spectral density $S_{NN}(f)$ consists of a sum of delta functions in accordance with the amplitudes and frequencies of the sinusoidal components of the triangular autocorrelation (71). This is in agreement with Eqs. (64), (65), and (67).

The noise performance evaluation of the MAP phase estimator for the square wave requires that we obtain the probability distribution of the value of x that maximizes Eq. (70) where $N_z(x)$ is gaussian with autocorrelation (71). It is clear that the unconditional density will be an even function of x about x_0. However, the conditional density (with uniform a priori x)

$$f_{x|z}(x|\vec{z}) = C_4(\vec{z}) \exp \frac{2}{N_0}\left[A_c T \operatorname{Cos} 2\pi(x - x_0) + N_z(x)\right] \tag{72}$$

is generally not an even function of x about x_0 and so the MAP estimator is not necessarily the MMSE, etc. estimator.

The small noise performance of the MAP phase estimator for the square wave is arbitrarily good in the sense that as a square wave is approximated by including more and more of its harmonics, the length of the closed path on the hypersphere surface increases without limit and the ds/dx of Eq. (52) approaches infinity. However, in this case "small noise" is noise whose component of RMS length $\sqrt{N_0/2}$ along $\vec{y}'(x_0)$ is small compared to the distance Δs along which the curve $\vec{y}(x)$ is approximately straight. This distance Δs is a fraction of the amplitude of the highest harmonic in $y(\cdot)$ and becomes arbitrarily small as the square wave is better and better approximated. Thus, the ideal square wave is a special case having no "linearized small noise" result for the MAP phase estimator. Some previous work (Ref. 3) on phase estimation of square waves is in agreement with this insight.

It appears that although square waves are easily generated and manipulated in electronic circuits, they are more difficult to deal with than sinusoids when estimating phase. This is generally true for any $\vec{y}(x)$ which spans a linear subspace of more than two dimensions. To find the value of x that maximizes $f_{x|z}(x|\vec{z})$, $\vec{z}$ must be observed in L_y and not in some linear subspace of smaller dimensions; that is, the number of observables must equal the dimension of L_y. If we use a smaller number of observables, we are dealing with an orthogonal projection of $\vec{y}(x)$ and $\vec{z}$ onto some linear subspace $\overline{L}_y$ of L_y. Then choosing x to minimize the distance between $\vec{y}(x)$ and $\vec{z}$ in $\overline{L}_y$ generally does not minimize the distance in L_y, and so the result is a suboptimal estimate. For the small noise case, the variance of the component of $\vec{n}$ along the projection of $\vec{y}'(x_0)$ onto $\overline{L}_y$ is the same as the variance of the component of $\vec{n}$ along $\vec{y}'(x_0)$ (in L_y).

But if $\bar{s}$ is the arclength of the projection onto $\overline{L}_y$, then generally

$$\frac{d\bar{s}}{dx} < \frac{ds}{dx} \tag{73}$$

and by Eq. (54) the variance of the error in $\hat{x}_0$ is increased.

For example, in the case of a square wave where the observables chosen are

$$\overline{I} = \int_{t_1}^{t_2} z(\tau) \operatorname{Cos} 2\pi\tau\, d\tau \tag{74a}$$

$$\overline{Q} = \int_{t_1}^{t_2} z(\tau) \operatorname{Sin} 2\pi\tau\, d\tau \tag{74b}$$

instead of Eq. (65), as in the idealization of the sequential ranging system, the above observations apply. L_y is infinite dimensional and $\overline{L}_y$ is two dimensional. Here we have $\vec{y}(x)$ given by

$$y(f_c t + x) = \frac{A_c}{\sqrt{2}} \operatorname{Cos} 2\pi(f_c t + x) \tag{75}$$

with

$$\vec{v}_1 = \frac{\vec{y}(0)}{\|\vec{y}\|} \tag{76a}$$

$$\vec{v}_2 = \frac{\vec{y}(-1/4)}{\|\vec{y}\|} \tag{76b}$$

as orthonormal basis vectors for $\bar{L}_y$. The small noise performance easily follows from

$$\left(\frac{ds}{dx}\right)^2 = \left(\frac{dy_1}{dx}\right)^2 + \left(\frac{dy_2}{dx}\right)^2 = 8A_c^2 T \sqrt{(\mathrm{Cos}^2 2\pi x + \mathrm{Sin}^2 2\pi x)}$$

$$= 16A_c^2 T, x \neq 0, \pm 1/4, 1/2 \tag{77}$$

Substituting this in the middle member of Eq. (54) gives

$$\sigma^2(\hat{x}_0) = \frac{N_0}{32A_c^2 T} \tag{78}$$

Thus the small noise performance for this case is 0.9 db worse than for a sinusoid of the same power and frequency (in the presence of the same noise).

The performance of the suboptimal estimator using the observables in Eq. (74) has been completely evaluated in Refs. 4, 5, and 6 including the singular behaviour at $x = 0$, $\pm 1/4$, $1/2$ in the small noise case. It is interesting to observe that optimum estimation of phase based on the incomplete set of observables in Eq. (74) requires a knowledge of the amplitude of the square wave in $z(t)$ – a piece of information that is not needed by the MAP estimator that maximizes Eqs. (69) or (70).

If we return to the full infinite dimensional basis for the square wave we see that the radius of the circle described by $\vec{y}(x)$ in the plane corresponding to each harmonic decreases inversely as the frequency of harmonic while the variance of the projection of $\vec{n}$ on any plane is N_0. Thus for a given value of $A_c^2 T/N_0$ there will be a highest harmonic in whose plane the signal to noise ratio is greater than unity. The observables corresponding to higher harmonics will be mostly noise and therefore should be more lightly weighted. This leads to the question of what is the useful number of observables to use as a function of $A_c^2 T/N_0$. An almost equivalent question is what is the useful number of harmonics in $y(\cdot)$ as a function of $A_c^2 T/N_0$. This has been in effect answered in the context of an indirect estimator in Ref. 3. There the optimum local reference waveform is essentially the derivative $\vec{y}'(x)$ of the $\vec{y}(x)$ having the useful number of harmonics.

The work reported above has benefited significantly from conversations with S. A. Butman and J. R. Lesh.

References

1. Rauch, L. L., *Improved Demodulation and Predetection Recording,* USAF Research Report, Contract No. AF-33(616)-5796; February 28, 1961.

2. Davenport, W. B., and Root, W. L., *An Introduction to the Theory of Random Signals and Noise,* Section 8-6, McGraw-Hill; 1958.

3. Layland, J. W., "On the Optimum Correlation Function for a First-Order Tracking Loop," *Space Program Summary 37-50,* Vol. III, pages 284-287. Jet Propulsion Laboratory, Pasadena, Calif., April 30, 1968.

4. Goldstein, R. M., "Ranging with Sequential Components", *Space Program Summary 37-52,* Vol II, pages 46-69. Jet Propulsion Laboratory, Pasadena, Calif., July 31, 1968.

5. Butman, S., "On Estimating the Phase of a Square Wave in White Noise," *Space Program Summary 37-53,* Vol. II, pages 200-209. Jet Propulsion Laboratory, Pasadena, Calif., October 31, 1968.

6. Butman, S., "The Effect of Amplitude Uncertainty on Estimating Phase of a Square Wave," *Space Program Summary 37-58.* Jet Propulsion Laboratory, Pasadena, Calif., August 31, 1969.